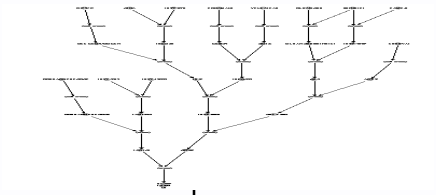
Enabling Biology to Become a Predictive Science

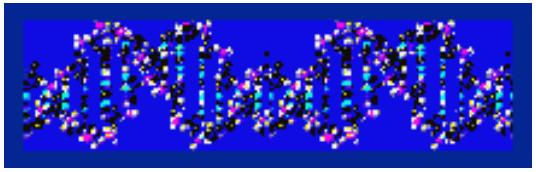
William D. Beavis, Geert Wenes, John Ambrosiano

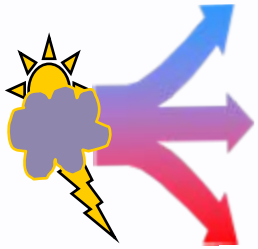
Computational Infrastructure for GTL Workshop
22 January 2002

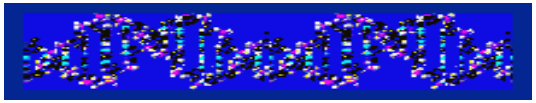
NATIONAL CENTER FOR GENOME

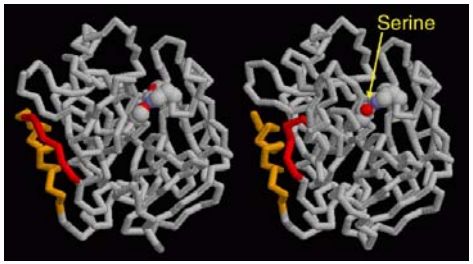
Goal: P(Phenotype|Data,Model)

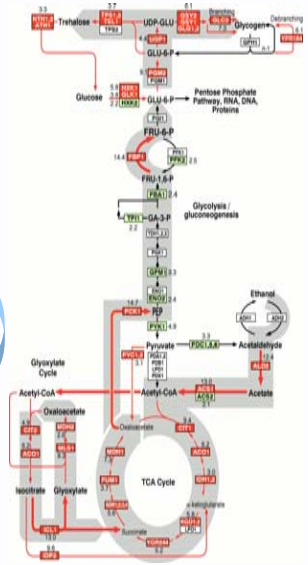












Discovery-Driven Data, Models and Computing Resources for Predicting the Phenotype.

	Data and Quality	Models and Lack of Fit	Compute Resources
Ancestry & Environment			
DNA			
*m*RNA			
Proteins			
Biochemical Networks			

Ancestry and Environment

Data:

- Pedigrees, Breeding Records, Progeny Performance Trials
- Family Histories, Clinical Records, Environmental Exposure
- 10^5 - 10^7 records per trait in geographically distributed, heterogenous repositories.
- $\varepsilon \sim .05$ - $.5$

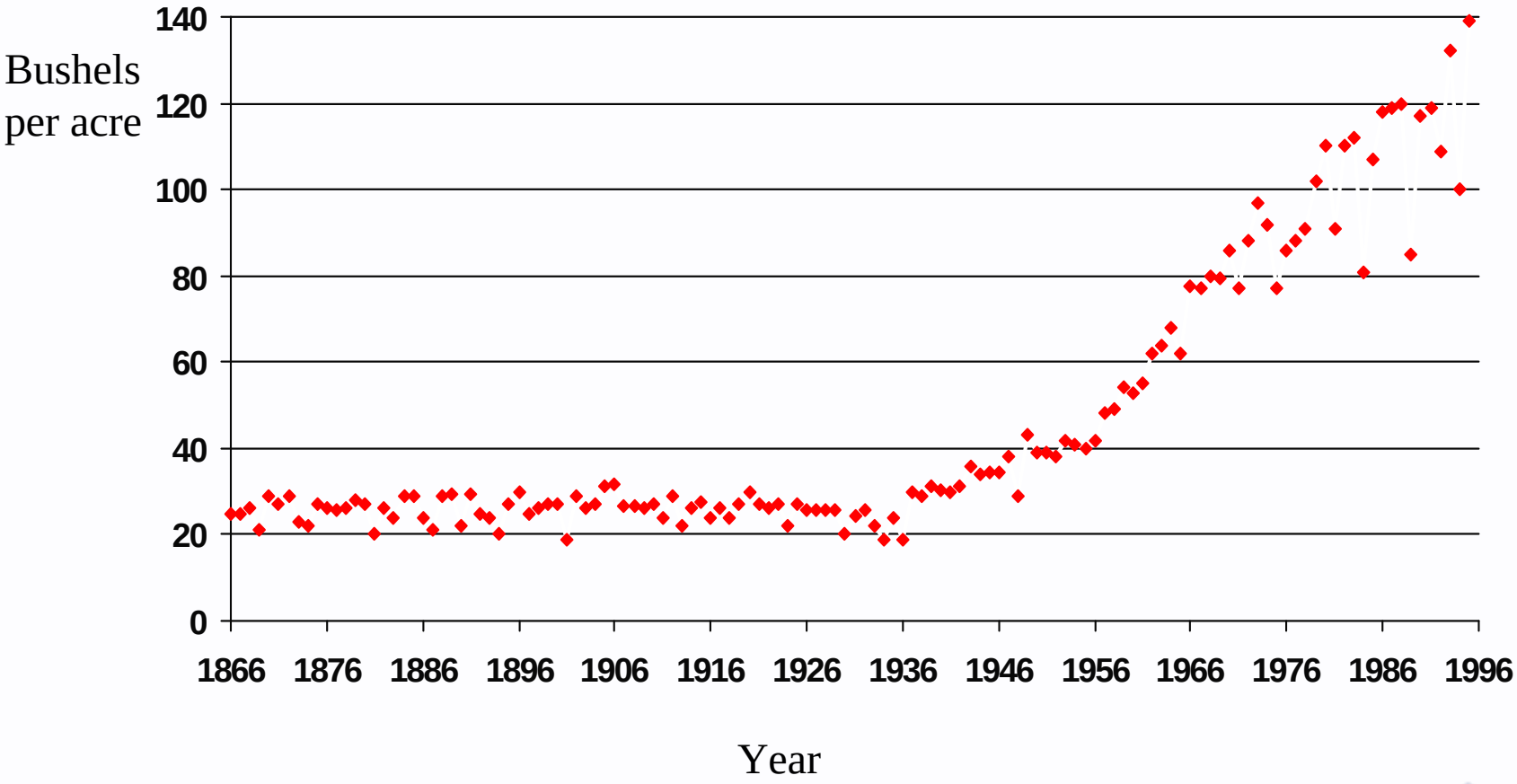
Models:

- $y = X\beta + Z\mu + \varepsilon$

Compute Resources:

- For brute-force ML, RAM needs to accommodate data from multiple sources and a few million records.

Average U.S. Corn Yields



Predictions from DNA information

Data:

- ⌘ Physical Maps, Linkage Maps and SNPs
- ⌘ Nucleotide Sequences
- ⌘ 1.6×10^{10} sequences in a few geographically distributed repositories.
- ⌘ Sequencing erros, $\varepsilon \sim .01$; annotation errors ?

Models:

- ⌘ $\mathbf{y = X\beta + Z\mu + W\iota + \varepsilon}$
- ⌘ **Blast (seq$_i$ ~ seq$_j$ => function) ;** $s_{ij} = \ln(q_{ij}/P_i P_j) / \lambda_u$

Compute Resources:

- ⌘ 200K ESTs: 2 weeks, 16-node Linux cluster with 16 Gb RAM

Weakness of sequence annotations:

Function assigned based on sequence similarity to another sequence with a

function assigned based on sequence similarity to another sequence with a

Function assigned based on sequence similarity to another sequence with a ……

Computerized stamp collecting

Predictions from DNA information

Data:

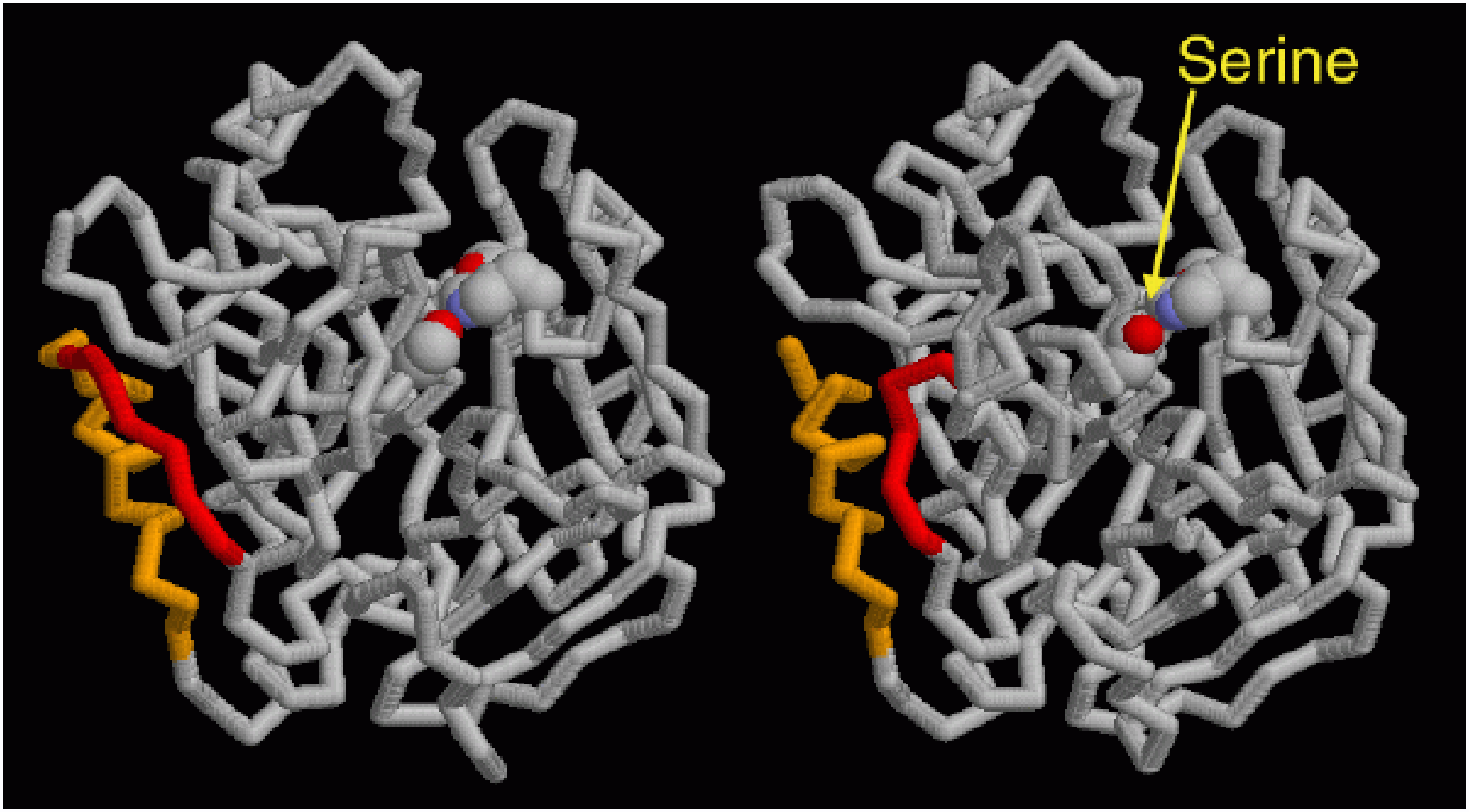
- ⌘ Physical Maps, Linkage Maps and SNPs
- ⌘ Nucleotide Sequences
- ⌘ 1.6×10^{10} sequences in a few geographically distributed repositories.
- ⌘ Sequencing erros, $\varepsilon \sim .01$; annotation errors ?

Models:

- ⌘ $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \mathbf{W}\iota + \varepsilon$
- ⌘ Blast ($seq_i \sim seq_j$ => function) ; $s_{ij} = \ln(q_{ij}/P_i P_j) / \lambda_u$
- ⌘ **CASP (seq => structure => function)**

Compute Resources:

- ⌘ Moderate sized clusters of commodity-grade processors.


Serine

Understanding the Mechanisms of Protein Folding

⌘ Will lead to an understanding of protein function in various cellular environments

(Misfolding is known to occur and be responsible for serious diseases)

⌘Resulting in development of predictable diagnostics, novel protein-based therapeutics, novel proteins for sequestering C, heavy metals, and radioactive isotopes.

Understanding the Mechanisms of Protein Folding

Experimental techniques are limited for relevant time scales, thus there is a need for simulation of

- ⌘ **folding kinetics**
- ⌘ **folding pathways**
- ⌘ **force-field assessments**

Allen et al. 2001. "Blue Gene: A vision for protein science using a petaflop supercomputer". IBM Systems Journal 40:310-327.

Levitt et al. 2002. "Modeling Across the Scales - Atoms to Organisms" Mathematics and Molecular Biology VII. Program in Mathematics and Molecular Biology. Santa Fe, N. Mex.

Computing Needs for Mechanisms of Protein Folding

Time frame to simulate	**10^{-4} sec**
Time-step size	**10^{-15} sec**
Number of MD time steps	**10^{11}**
Atoms in a typical protein/water simu	**3.2×10^{5}**
Number of interactions in a force calc	**10^{9}**
Instructions per force calc	**10^{3}**
Total number of instructions	**10^{23}**

Allen et al. 2001. "Blue Gene: A vision for protein science using a petaflop supercomputer" IBM Systems Journal 40:310-327.

Predictions from mRNA and proteomic arrays

Data:

- ⌘ gene-chips, micro-arrays, bead-arrays (MPSS), proteomic arrays
- ⌘ very sparse data, relative to biological time scales, snap shots of average transcription from pools of similarly treated cells
- ⌘ 10^4 - 10^5 genes per experiment; 10^4 experiments in geographically distributed, heterogenous repositories.
- ⌘ $\varepsilon \sim .1$ - $.5$

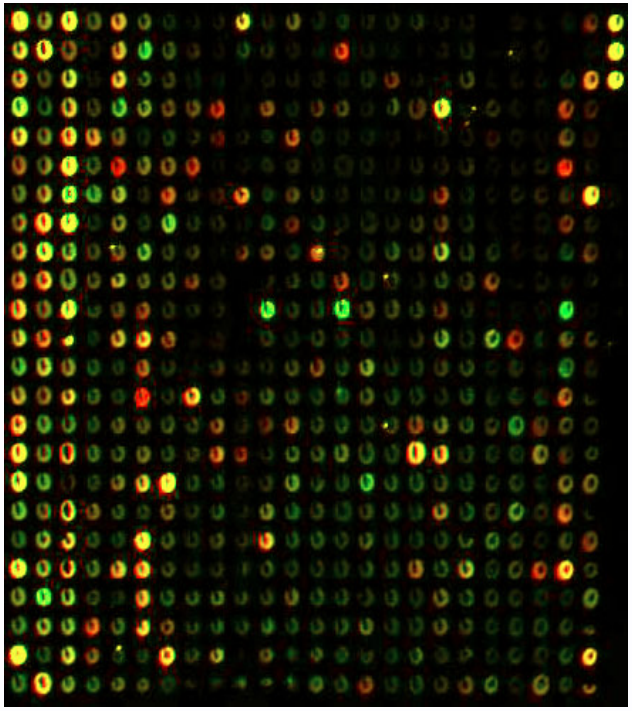
Models:

- ⌘ $y = X\beta + Z\mu + W\iota + T\nu + \varepsilon$
- ⌘ **Support Vector Machines**

Compute Resources:

- ⌘ RAM needs to accommodate data from multiple arrays, multiple experiments, GenBank, SwissProt, Kegg, Pfam, PathDB

Examples of data from micro-arrays and 2 D gel protein arrays



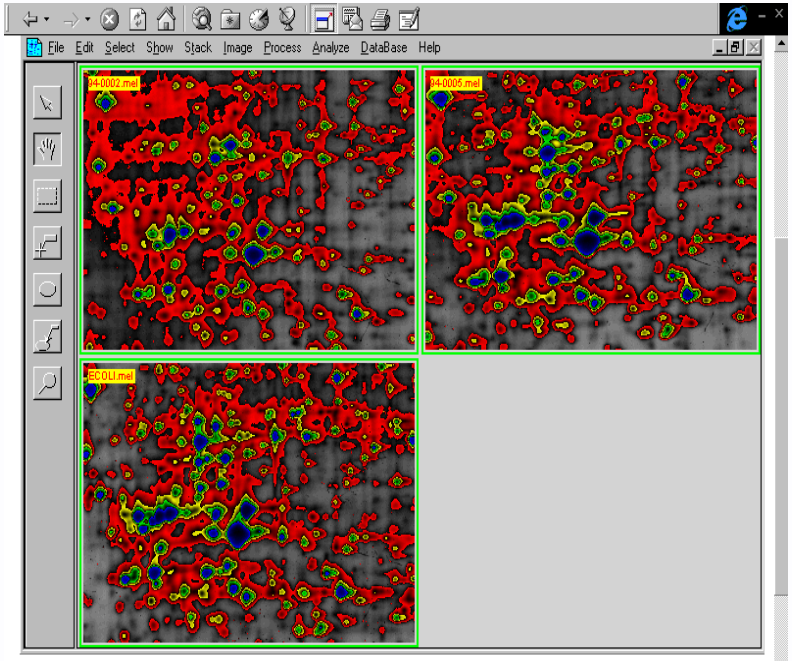
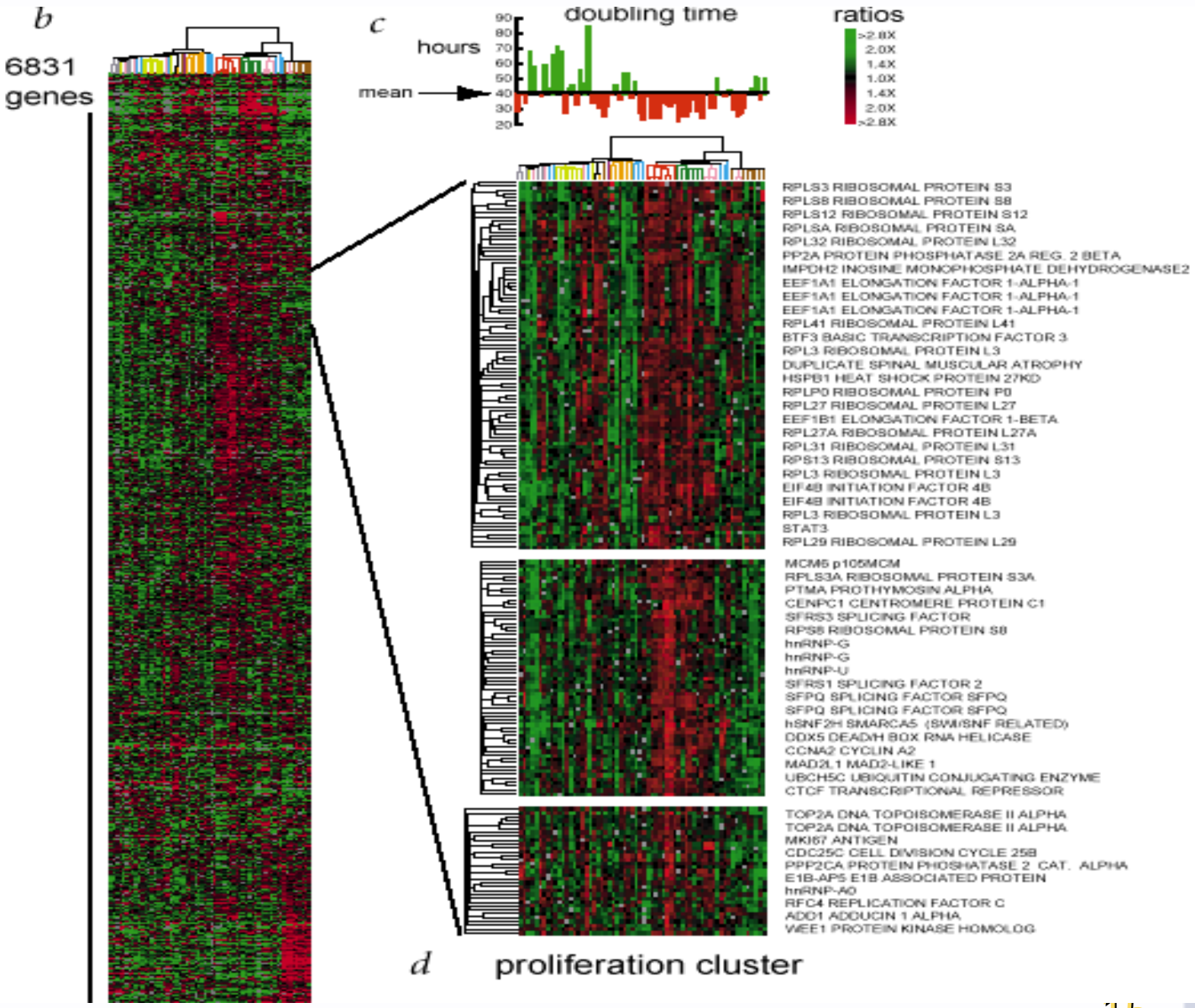


Figure 17-8. Gels displayed using the *FiveRamps* color lookup table.


b
6831
genes
c
hours
mean
doubling time
ratios
>2.8X
2.0X
1.4X
1.0X
1.4X
2.0X
>2.8X
RPLS3 RIBOSOMAL PROTEIN S3
RPLS8 RIBOSOMAL PROTEIN S8
RPLS12 RIBOSOMAL PROTEIN S12
RPLSA RIBOSOMAL PROTEIN SA
RPL32 RIBOSOMAL PROTEIN L32
PP2A PROTEIN PHOSPHATASE 2A REG. 2 BETA
IMPDH2 INOSINE MONOPHOSPHATE DEHYDROGENASE2
EEF1A1 ELONGATION FACTOR 1-ALPHA-1
EEF1A1 ELONGATION FACTOR 1-ALPHA-1
EEF1A1 ELONGATION FACTOR 1-ALPHA-1
RPL41 RIBOSOMAL PROTEIN L41
BTF3 BASIC TRANSCRIPTION FACTOR 3
RPL3 RIBOSOMAL PROTEIN L3
DUPLICATE SPINAL MUSCULAR ATROPHY
HSPB1 HEAT SHOCK PROTEIN 27KD
RPLP0 RIBOSOMAL PROTEIN P0
RPL27 RIBOSOMAL PROTEIN L27
EEF1B1 ELONGATION FACTOR 1-BETA
RPL27A RIBOSOMAL PROTEIN L27A
RPL31 RIBOSOMAL PROTEIN L31
RPS13 RIBOSOMAL PROTEIN S13
RPL3 RIBOSOMAL PROTEIN L3
EIF4B INITIATION FACTOR 4B
EIF4B INITIATION FACTOR 4B
RPL3 RIBOSOMAL PROTEIN L3
STAT3
RPL29 RIBOSOMAL PROTEIN L29
MCM6 p105MCM
RPLS3A RIBOSOMAL PROTEIN S3A
PTMA PROTHYMOSIN ALPHA
CENPC1 CENTROMERE PROTEIN C1
SFRS3 SPLICING FACTOR
RPS8 RIBOSOMAL PROTEIN S8
hnRNP-G
hnRNP-G
hnRNP-U
SFRS1 SPLICING FACTOR 2
SFPQ SPLICING FACTOR SFPQ
SFPQ SPLICING FACTOR SFPQ
hSNF2H SMARCA5 (SWI/SNF RELATED)
DDX5 DEAD/H BOX RNA HELICASE
CCNA2 CYCLIN A2
MAD2L1 MAD2-LIKE 1
UBCH5C UBIQUITIN CONJUGATING ENZYME
CTCF TRANSCRIPTIONAL REPRESSOR
TOP2A DNA TOPOISOMERASE II ALPHA
TOP2A DNA TOPOISOMERASE II ALPHA
MKI67 ANTIGEN
CDC25C CELL DIVISION CYCLE 25B
PPP2CA PROTEIN PHOSPHATASE 2 CAT. ALPHA
E1B-AP5 E1B ASSOCIATED PROTEIN
hnRNP-A0
RFC4 REPLICATION FACTOR C
ADD1 ADDUCIN 1 ALPHA
WEE1 PROTEIN KINASE HOMOLOG
d
proliferation cluster

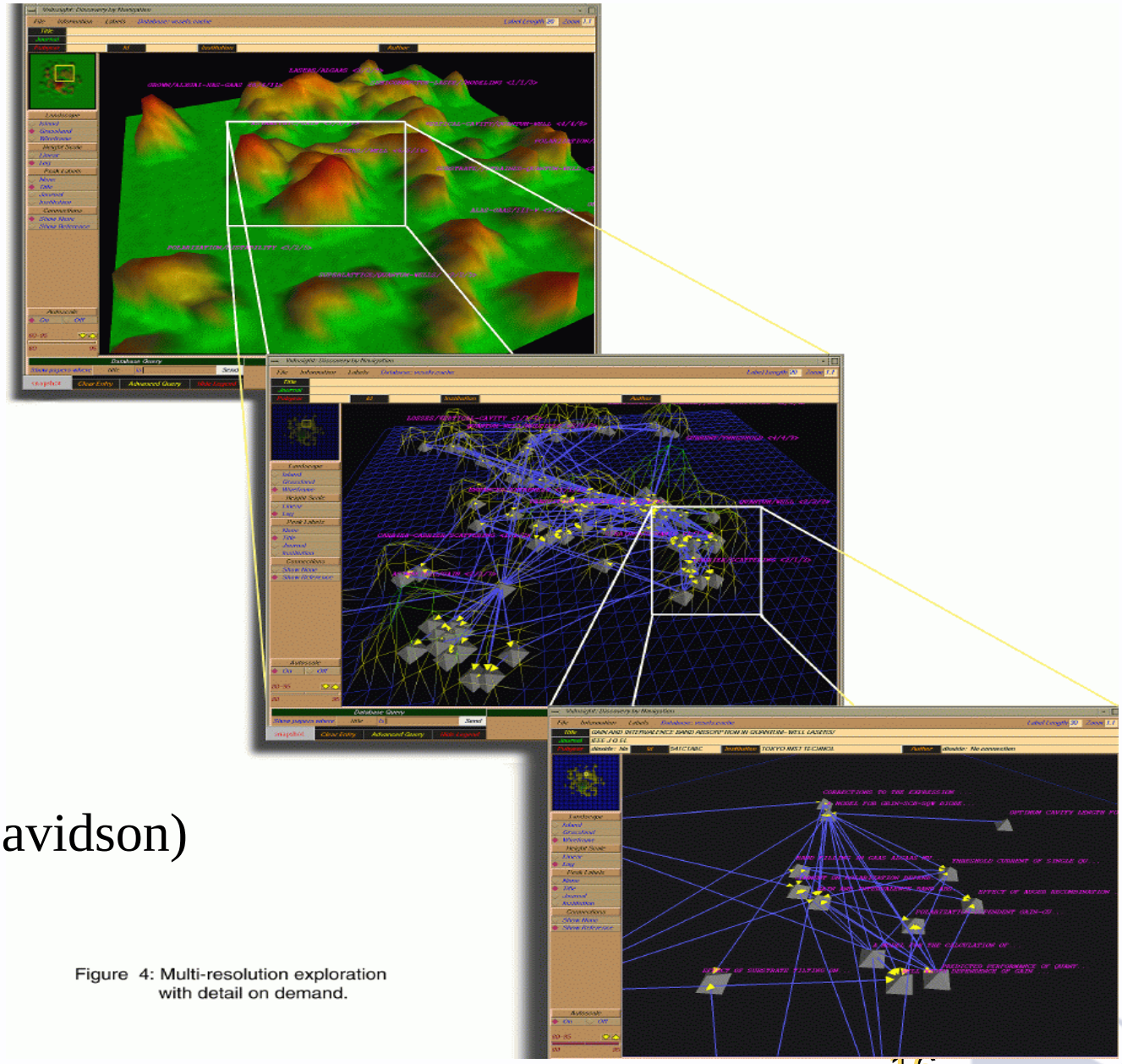


Figure 4: Multi-resolution exploration with detail on demand.

VxInsight
(SNL: George Davidson)

Predictions from Expression Arrays

Data:

- ⌘ gene-chips, micro-arrays, bead-arrays (MPSS), proteomic arrays
- ⌘ very sparse data, relative to biological time scales, snap shots of average transcription from pools of similarly treated cells
- ⌘ 10^4 - 10^5 genes per experiment; 10^4 experiments in geographically distributed, heterogenous repositories.
- ⌘ $\varepsilon \sim .1$ - $.5$

Models:

- ⌘ $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \mathbf{W}\iota + \mathrm{T}\nu + \varepsilon$
- ⌘ **Support Vector Machines**

Compute Resources:

- ⌘ RAM needs to accommodate data from multiple arrays, multiple experiments, GenBank, SwissProt, Kegg, Pfam, PathDB

Pathways and Networks

Data:

- expression arrays
- cross-link arrays of protein-protein, protein-DNA and protein-RNA interactions.
- Very sparse data, i.e., snap shots at discrete time/treatment/tissue samples
- geographically distributed, hetoerogenous repositories

Models:

- Hypothetical networks of nodes and edges, e.g., Baysian Networks

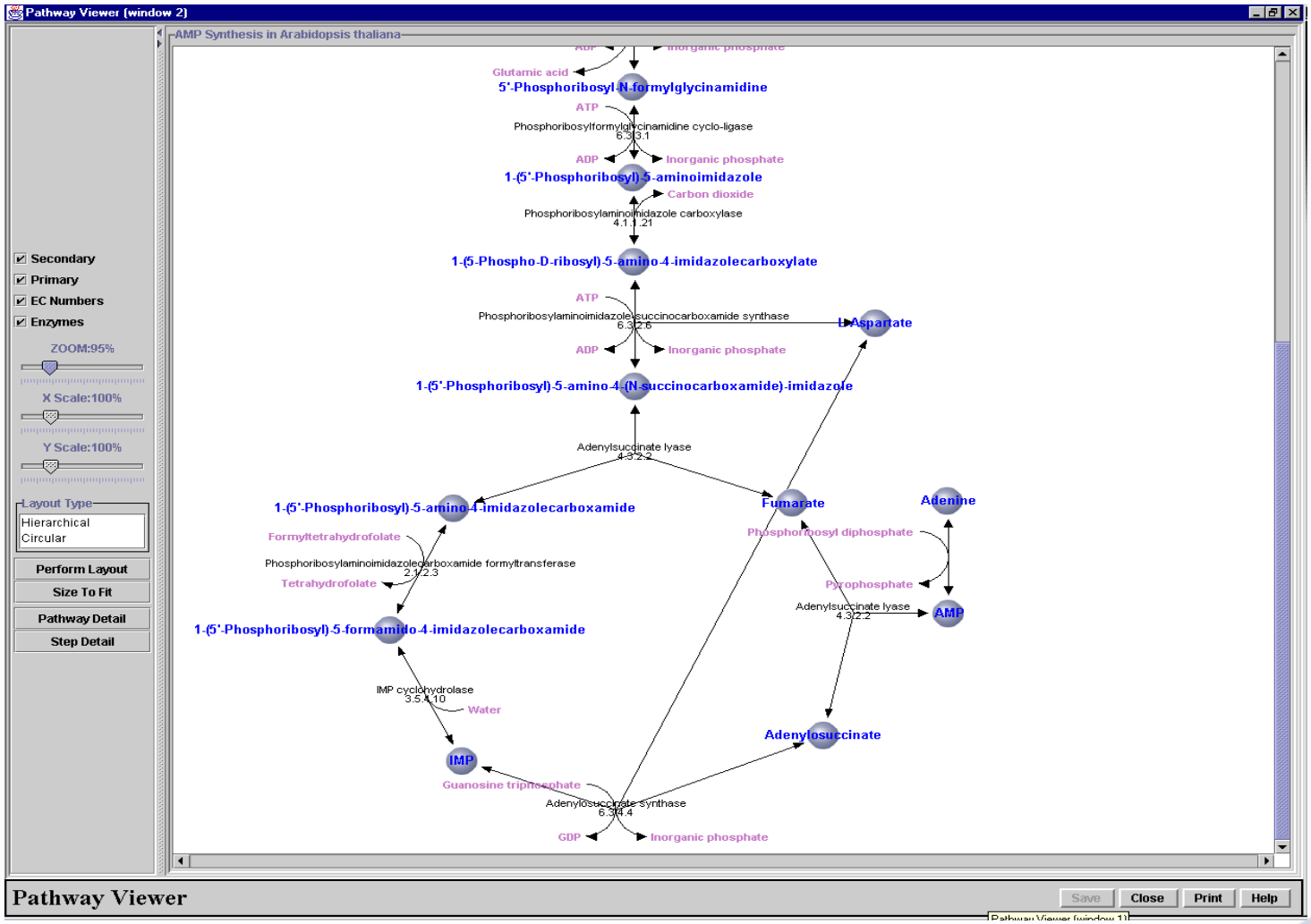
Compute Resources:

- Similar to “radiation transport” simulations?

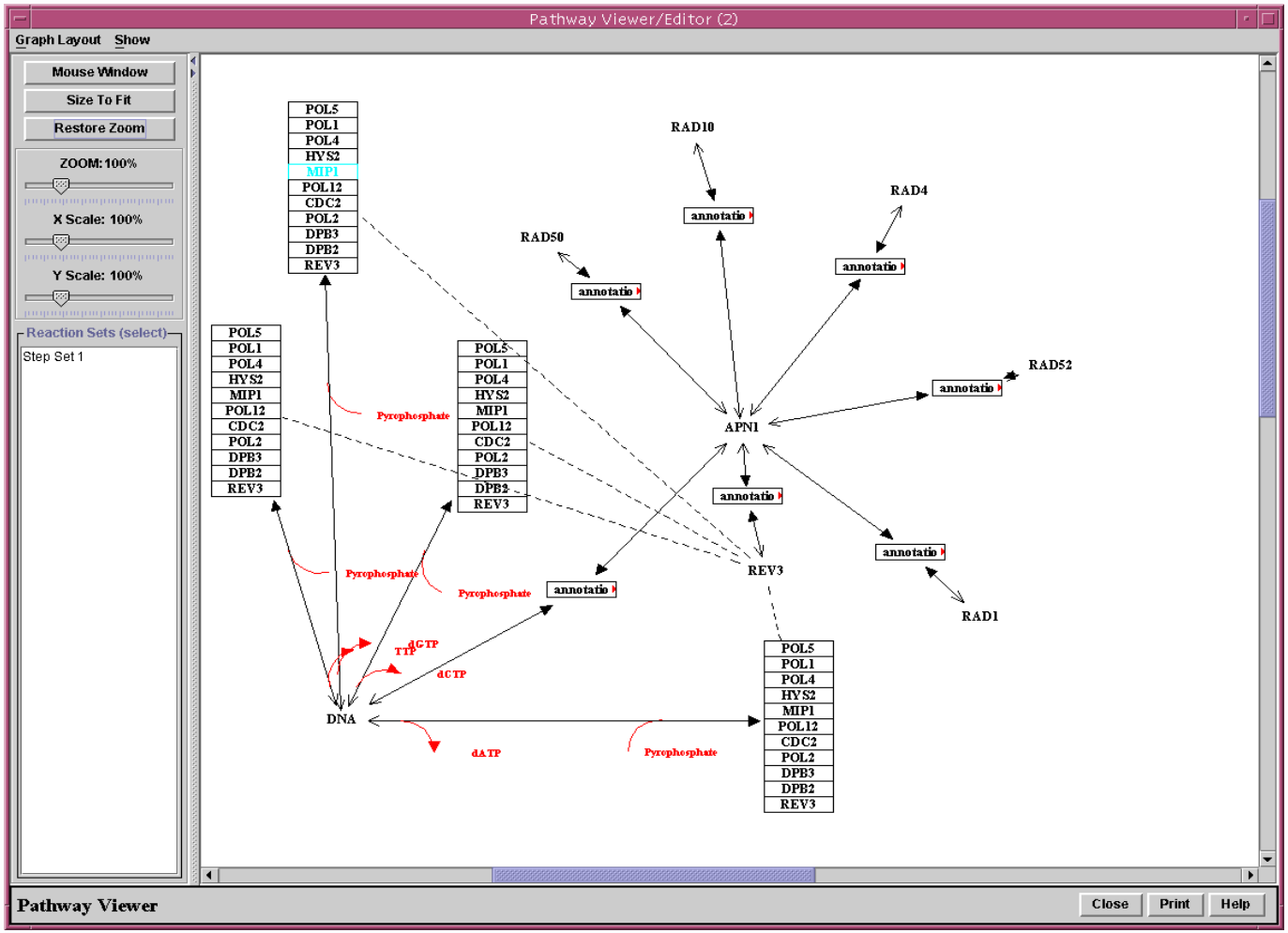
Data Needed for Inferring Genetic Networks from N Genes

Boolean, Fully Connected	2^N
Boolean, K Connections per Gene	$2^K[\log(N)]$
Boolean, K Connections per gene based on linearly separable functions	$K[\log(N)]$

D'haeseleer, 1997. Data Requirements for Gene Network Inference
http://www.cs.unm.edu/~patrik/networks/


Pathway Viewer (window 2)
AMP Synthesis in Arabidopsis thaliana
Glutamic acid
5'-Phosphoribosyl-N-formylglycinamidine
ATP
Phosphoribosylformylglycinamidine cyclo-ligase
6.3.3.1
ADP
Inorganic phosphate
1-(5'-Phosphoribosyl)-5-aminoimidazole
Carbon dioxide
Phosphoribosylaminoimidazole carboxylase
4.1.1.21
1-(5-Phospho-D-ribosyl)-5-amino-4-imidazolecarboxylate
ATP
Phosphoribosylaminoimidazolesuccinocarboxamide synthase
6.3.2.6
L-Aspartate
ADP
Inorganic phosphate
1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole
Adenylsuccinate lyase
4.3.2.2
1-(5'-Phosphoribosyl)-5-amino-4-imidazolecarboxamide
Fumarate
Adenine
Formyltetrahydrofolate
Phosphoribosyl diphosphate
Phosphoribosylaminoimidazolecarboxamide formyltransferase
2.1.2.3
Tetrahydrofolate
Pyrophosphate
Adenylsuccinate lyase
4.3.2.2
AMP
1-(5'-Phosphoribosyl)-5-formamido-4-imidazolecarboxamide
IMP cyclohydrolase
3.5.4.10
Water
Adenylosuccinate
IMP
Guanosine triphosphate
Adenylosuccinate synthase
6.3.4.4
GDP
Inorganic phosphate
Secondary
Primary
EC Numbers
Enzymes
ZOOM:95%
X Scale:100%
Y Scale:100%
Layout Type
Hierarchical
Circular
Perform Layout
Size To Fit
Pathway Detail
Step Detail
Pathway Viewer
Save
Close
Print
Help

Simple network diagram including protein complexes, metabolic reactions and protein-protein interactions



Pathway Viewer/Editor

Graph Layout Show

Mouse Window

Size To Fit

Restore Zoom

ZOOM:46%

X Scale:100%

Y Scale:100%

Reaction Sets (select)

- New PathDB Look
- Alkaloid biosynthesis II - F
- Alkaloid biosynthesis I - F
- Flavonoids, stilbene and
- Xenobiotics metabolism -
- Sulfur metabolism - Refe
- Nitrogen metabolism - Re
- Terpenoid biosynthesis -
- Porphyrin and chlorophyll
- Retinol metabolism - Ref
- Folate biosynthesis - Ref
- Biotin metabolism - Refer
- Pantothenate and CoA bi
- Nicotinate and nicotinami
- Vitamin B6 metabolism -
- Riboflavin metabolism - F
- Thiamine metabolism - R
- Reductive carboxylate cyc
- Carbon fixation - Referen
- Methane metabolism - Re
- One carbon pool by folate
- C5-Branched dibasic acid

Choose a Color

Swatches HSB RGB

Recent:

Preview

Sample Text Sample Text

Sample Text Sample Text

Sample Text Sample Text

OK Cancel Reset

Record Navigator

Record: 85 of 107 85 - GD_21689.class

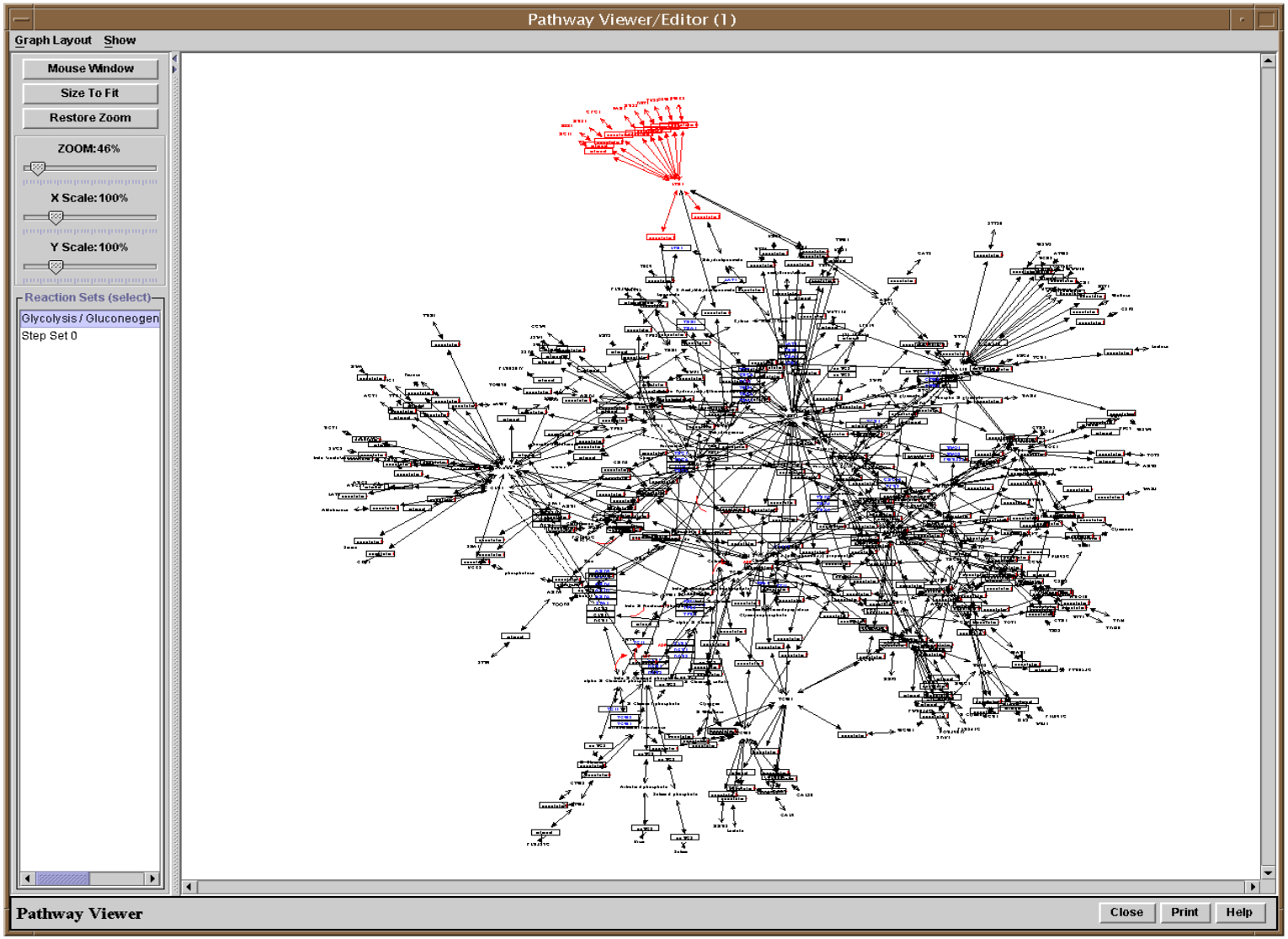
Pathway Viewer

Close Print Help

Office

Microsoft

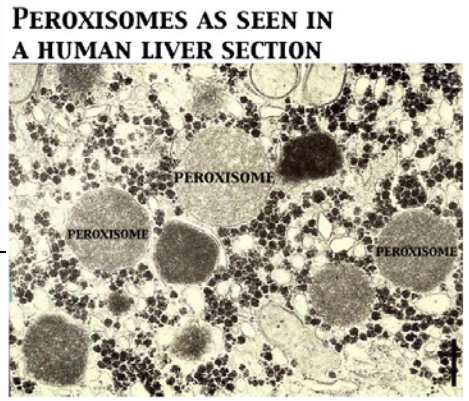
Start | Meeting Maker® - ... | Inbox - Outlook Ex... | Microsoft Word - ... | Microsoft PowerP... | My Computer | Find: Files named ... | C:\program file\Pa... | C:\WINNT\Syste... | Pathway View... | 1:05 PM

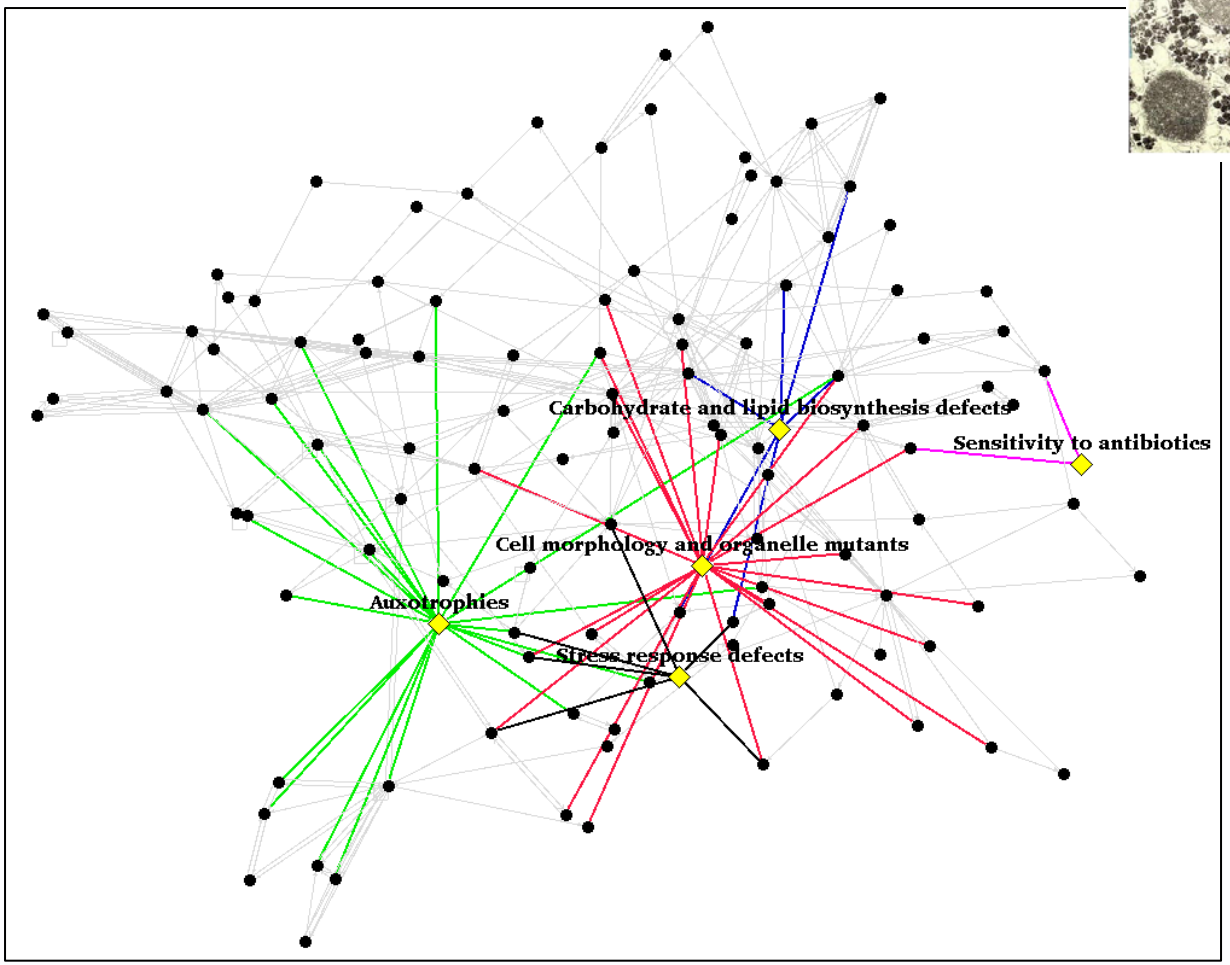

Pathway Viewer/Editor (1)
Graph Layout
Show
Mouse Window
Size To Fit
Restore Zoom
ZOOM: 46%
X Scale: 100%
Y Scale: 100%
Reaction Sets (select)
Glycolysis / Gluconeogen
Step Set 0
Pathway Viewer
Close
Print
Help

A Need for Integration

- The data are located in hundreds of geographically distributed, heterogenous repositories.
- Analysis tools (implemented algorithms) are likewise widely dispersed.
- Novel algorithms will be developed redundantly by independent PI.
- High performance computing resources are limited, but geographically distributed.

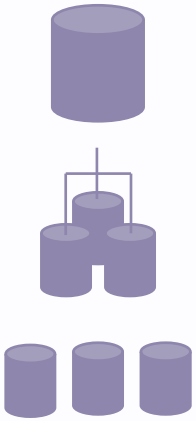
Functional information highlighted on a network of yeast peroxisomal proteins





Integration Taxonomy

Data Oriented



- Data Warehouse
- Federated Databases
- Multiple Databases

Software Oriented

- Web based (hyperlinks)
- Component based
  - proprietary
  - enterprise

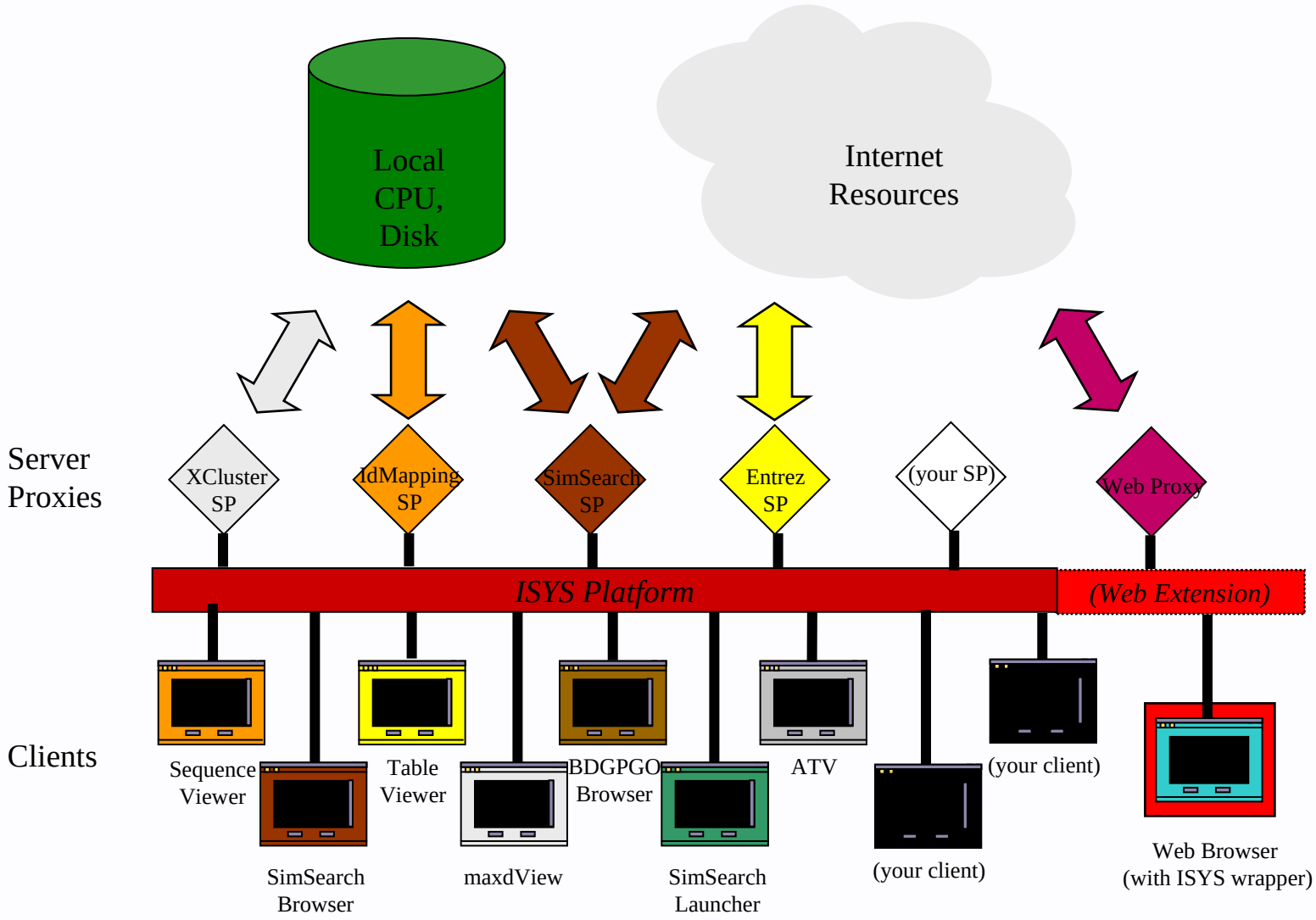
Complex database and software

- ⌘**Rigidity**
- ⌘**Fragility**
- ⌘**Expense**

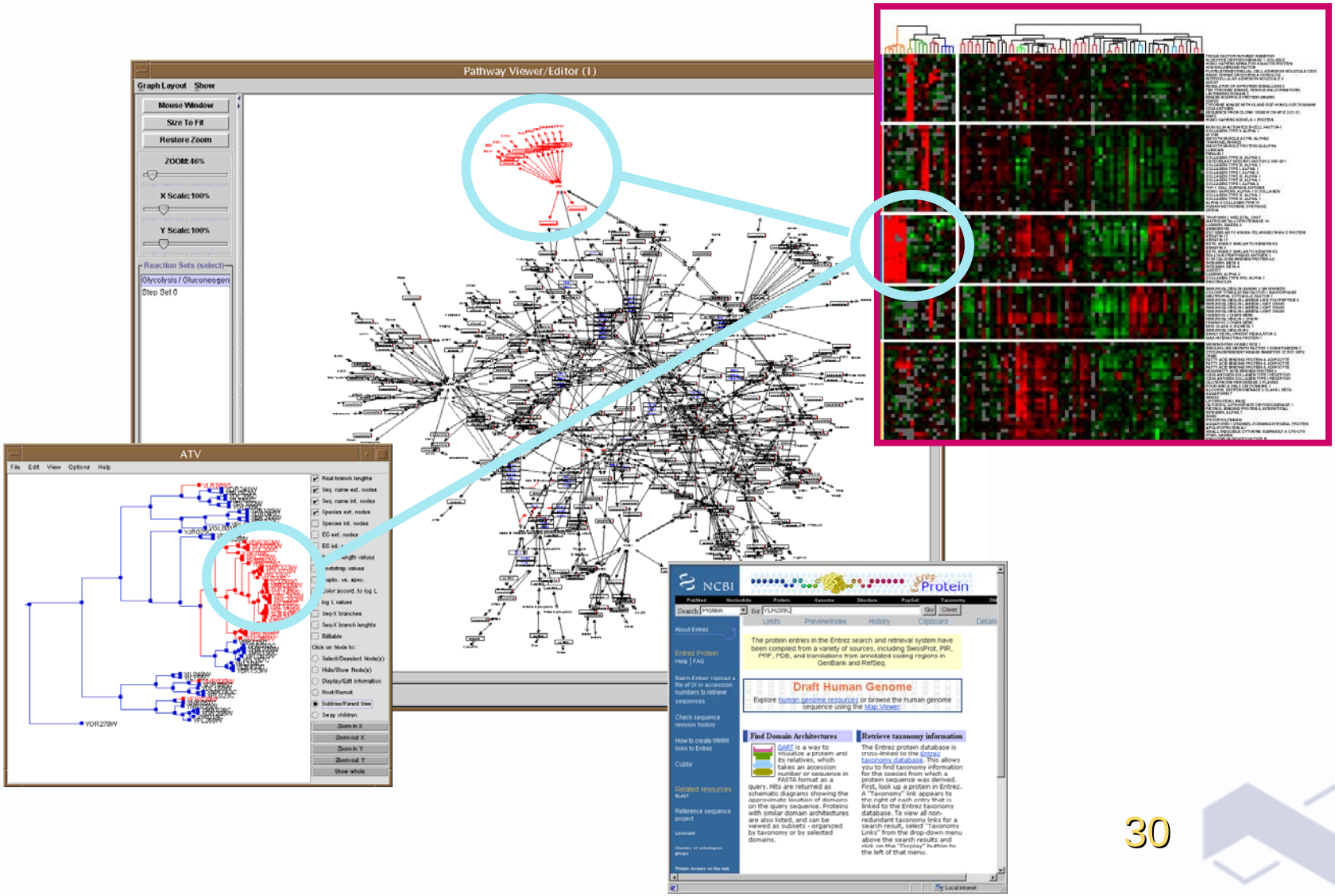
Software invariably lasts longer than you think (Y2K)

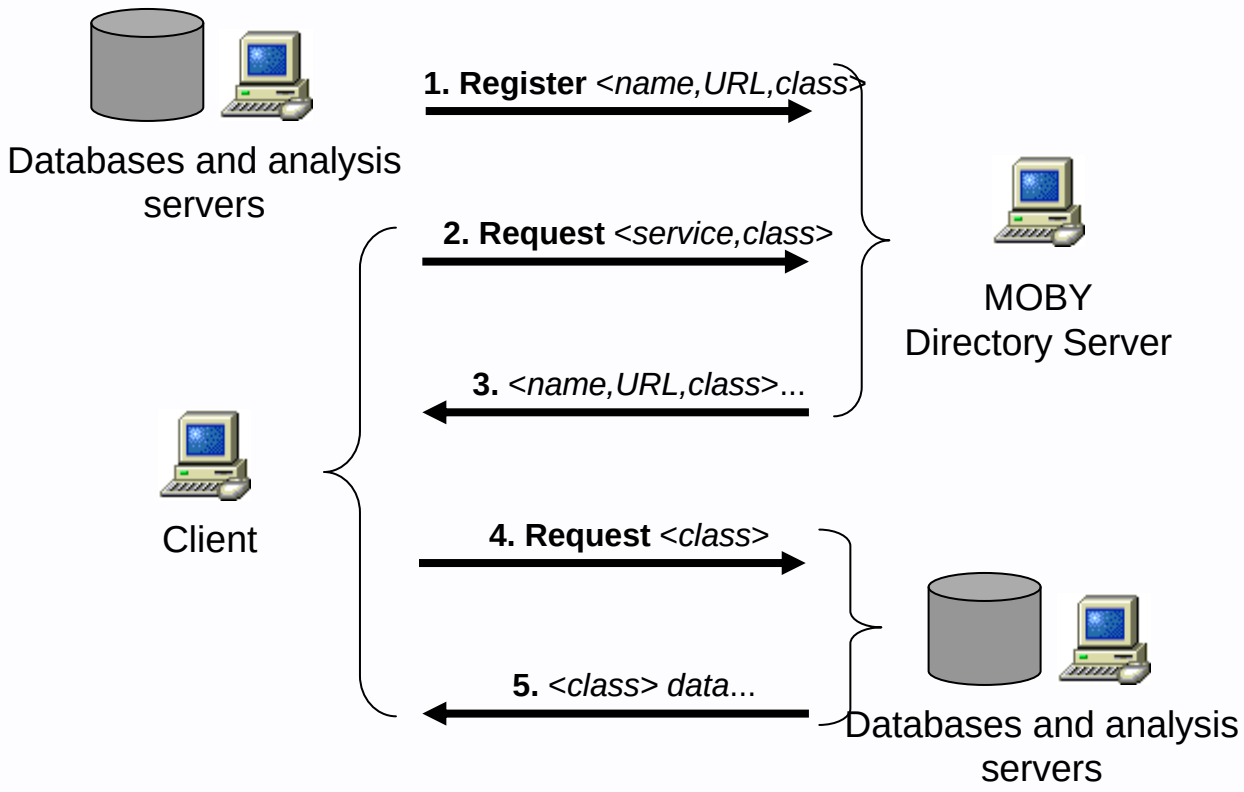
Loose Coupling Architecture

- ⌘Tight packaging of data and application allows context sensitive display of information.
- ⌘Plug and play architecture emphasizes communication between components via fundamental biological concepts.
- ⌘Separate components allow parallel and independent scientific advancements.


Local
CPU,
Disk
Internet
Resources
Server
Proxies
XCluster
SP
IdMapping
SP
SimSearch
SP
Entrez
SP
(your SP)
Web Proxy
ISYS Platform
(Web Extension)
Clients
Sequence
Viewer
Table
Viewer
BDGPGO
Browser
ATV
(your client)
SimSearch
Browser
maxdView
SimSearch
Launcher
(your client)
Web Browser
(with ISYS wrapper)

Integrated biological data types with an Integration Platform.




Databases and analysis servers
1. Register <name,URL,class>
2. Request <service,class>
MOBY Directory Server
3. <name,URL,class>...
Client
4. Request <class>
5. <class> data...
Databases and analysis servers

Discovery-Driven Data, Models and Computing Resources for Predicting the Phenotype.

	Data and Quality	Models and Lack of Fit	Compute Resources
Ancestry & Environment	Large volume Highly distributed low error	Linear Models Very large LOF	Commodity Clusters
DNA	Huge volume Few repositories very low error	Linear Models with a large LOF Annotation with potentially huge LOF	Commodity Clusters
*m*RNA	Huge volume Thousands of repositories Multiple Platforms high error	Linear Models with a LOF SVM ?	Large RAM
Proteins	Huge volume Hundreds of repositories Multiple Platforms high error	Protein Structure Predictions Protein Folding Mechanisms via simulation modeling?	Commercial SPM Blue-Gene +
Biochemical Networks	Very little experimental data	Baysian Networks Radiation Transport and other simulation models?	Commercial SPM Blue-Gene +

Enabling Biology to Become a Predictive Science

NCGR
Damian Gessler
Robert Kueffner
Honghui Wan
Jeff Blanchard
Mark Waugh

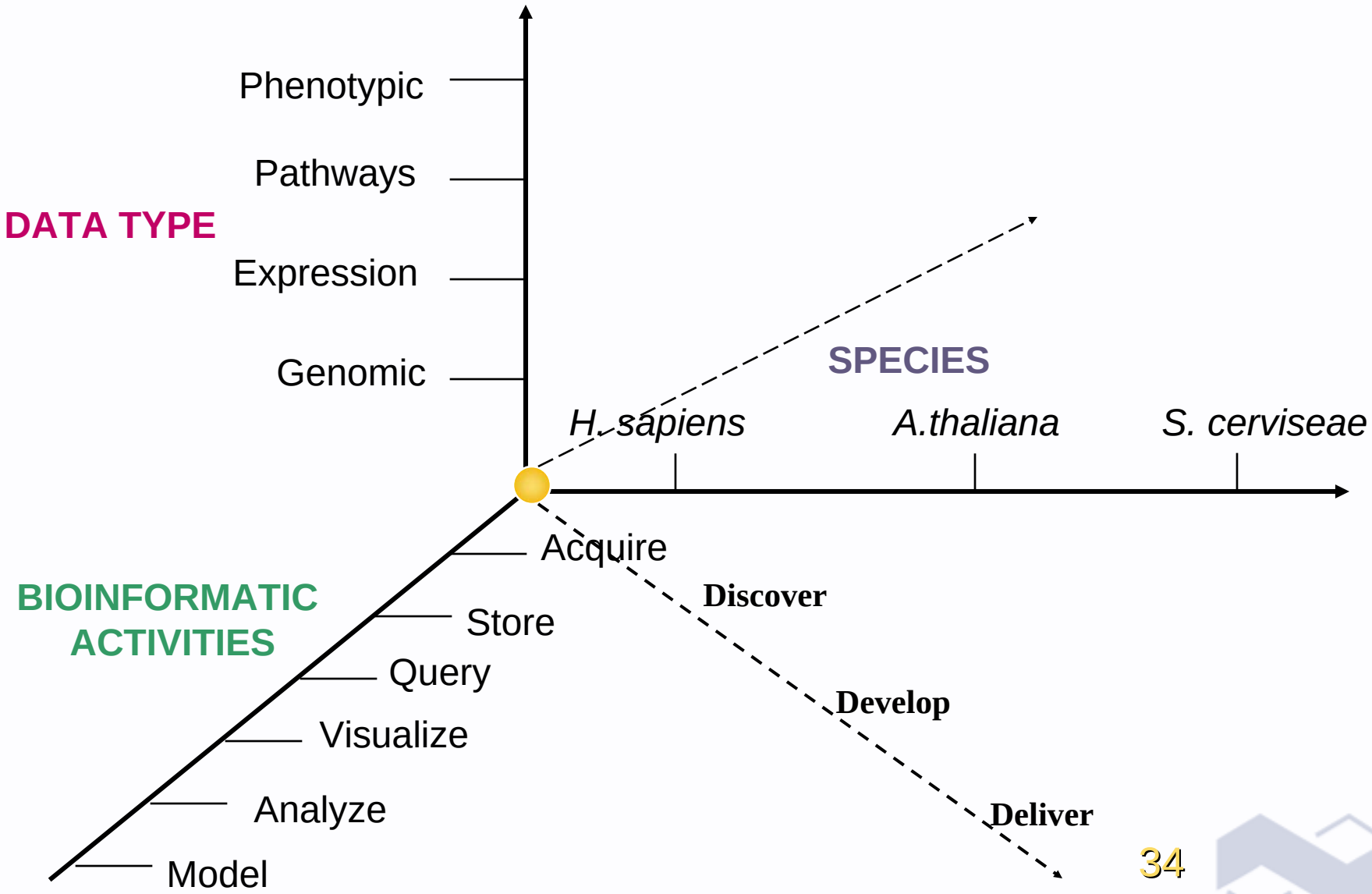
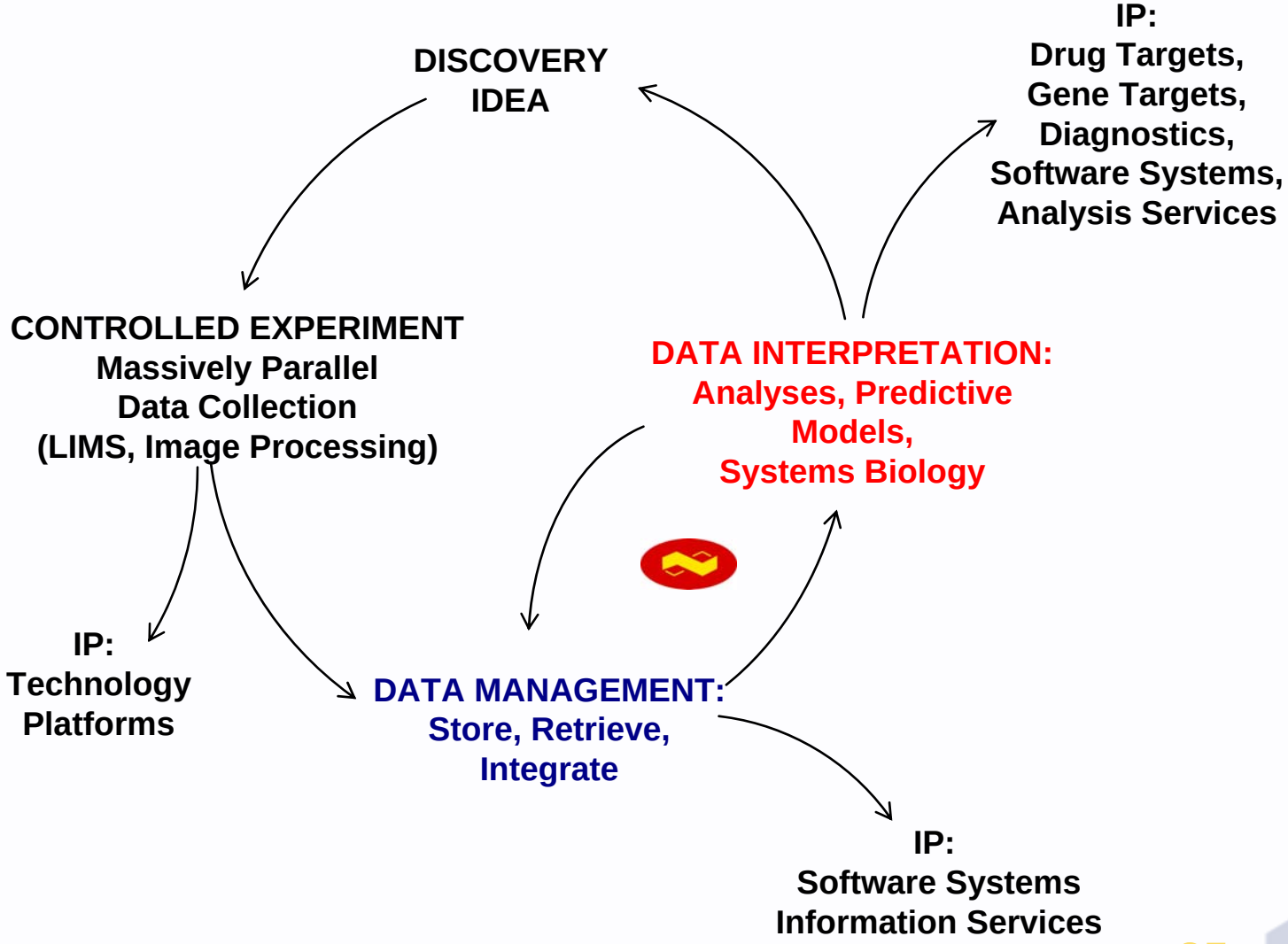
Sandia National Labs
Grant Hefflefinger
George Davidson

Joint Genome Institute
Dan Rokshar

Oak Ridge National Labs
Jerry Tuskan
Frank Larimer

Los Alamos National Labs
Mark Wall

Bioinformatics



DISCOVERY-DRIVEN BIOLOGICAL RESEARCH